AD-A051 090 STATE UNIV OF NEW YORK AT BUFFALO AMHERST STATISTICA--ETC F/G 12/1
NONPARAMETRIC STATISTICAL DATA SCIENCE: A UNIFIED APPROACH BASE--ETC(U)
JAN 77 E PARZEN DAAG29-76-G-0239
UNCLASSIFIED TR-47 ARO-13845.1-M NL

| OF |
ADA
051090

END
DATE
FILMED
4 -78
DDC

State University of New York at Buffalo
Department of Computer Science

# STATISTICAL SCIENCE

4230 Ridge Lea Road, Room A33; Amherst, New York 14226 Telephone: 716-- 831-1232



ARO 13845.1-M

(3)

NONPARAMETRIC STATISTICAL DATA SCIENCE:

A UNIFIED APPROACH BASED ON
DENSITY ESTIMATION AND
TESTING FOR "WHITE NOISE"*

by

Emanuel Parzen

Statistical Science Division
State University of New York at Buffalo





GRANT TECHNICAL REPORT NO. ARO-1
STATISTICAL SCIENCE DIVISION REPORT NO. 47

January 1977

*Research supported by the Army Research Office (Grant DA AG29-76-0239).

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report No. ARO-1
2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER

4. TITLE *(and Subtitle)*: Nonparametric Statistical Data Science: A Unified Approach Based on Density Estimation and Testing for "White Noise".

5. TYPE OF REPORT & PERIOD COVERED: Technical rept.

6. PERFORMING ORG. REPORT NUMBER: TR-47

7. AUTHOR(s): Emanuel Parzen

8. CONTRACT OR GRANT NUMBER(s): DAAG29-76-G-0239

9. PERFORMING ORGANIZATION NAME AND ADDRESS:
Statistical Laboratory
State University of New York at Buffalo
Amherst, New York 14226

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: ARO 13845.1-M

11. CONTROLLING OFFICE NAME AND ADDRESS

12. REPORT DATE: January 1977

13. NUMBER OF PAGES: 60 (63p.)

14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*

15. SECURITY CLASS. *(of this report)*: Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.



17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

NA

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

| | |
|---|---|
| Statistical Science | Nonparametric Regression |
| Nonparametric Statistics | Tests for Independence |
| Time Series Analysis | Goodness of Fit |
| Density Estimation | Order Statistics |
| Reliability | Rank Statistics |

Parameter estimation

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

The aim of this paper ~~is to~~ propose an approach to non-parametric statistical continuous data science which seems to be consistent with the conventional theories and methods of non-parametric inference but seems to point the way to universally applicable procedures (for continuous data) which are asymptotically as efficient as the best conventional goodness of fit and parameter estimation procedures available for each particular problem. ~~We have programmed~~ the methods described and

(continued on separate page)

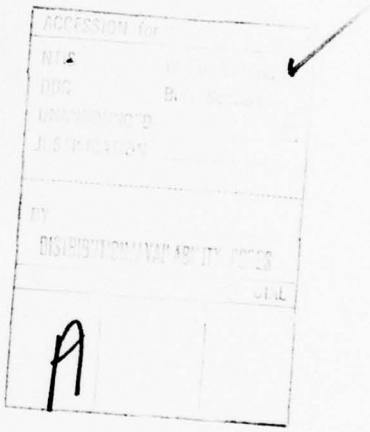
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

409 511

20. ABSTRACT (continued)

found them successful in test cases. However, in the space available to this paper we are only able to discuss (without proofs or examples) "Chapter 1" of our work which outlines the "ideas": how the basic general applied problems of statistical inference can be formulated as problems of estimation of distribution functions on the unit interval (or the unit hyper-cube), how such problems are more fruitfully treated as density estimation problems, and how to solve density estimation problems one can use the method which is the essence of the highly successful maximum likelihood method of parameter estimation: using a suitable information-theoretic divergence distance between densities, find the "smooth" density which is closest to a "raw" estimator of the density.



# NONPARAMETRIC STATISTICAL DATA SCIENCE: A UNIFIED APPROACH BASED ON DENSITY ESTIMATION AND TESTING FOR "WHITE NOISE"*

by

Emanuel Parzen

Statistical Science Division
State University of New York at Buffalo
Amherst, New York 14226

## Abstract

We demonstrate how many basic statistical inference problems (including the non-parametric one sample and multi-sample univariate and multivariate inference problems as well as time series problems) can be formulated as a hypothesis that a suitable distribution function $D(u)$, $0 \leq u \leq 1$ satisfies $D(u) = u$, $0 \leq u \leq 1$.

From the data one can construct a raw estimator $\tilde{D}(u)$ of $D(u)$, which has the property that asymptotically (as the sample size tends to $\infty$), under the null hypothesis that $D(u) = u$, $\sqrt{n}\{\tilde{D}(u) - u\}$, $0 \leq u \leq 1$, is a Brownian bridge stochastic process. A conventional statistical approach would be: test the hypothesis $D(u) = u$ by examining the significance of the deviation from zero of various functionals of $\tilde{D}(u) - u$.

The time series theoretic approach is to consider the density $d(u) = D'(u)$, $0 \leq u \leq 1$, and the Fourier Stieltjes transform

*Research supported by the Army Research Office (Grant DA AG29-76-G-0239) and by the State University of New York at Buffalo (sabbatical leave).

$$\varphi(v) = \int_0^1 e^{2\pi i u v} dD(u) , \qquad v = 0, \pm 1, \ldots$$

and to estimate them. Raw estimators are given by $\tilde{d}(u) = \tilde{D}'(u)$ , and

$$\tilde{\varphi}(v) = \int_0^1 e^{2\pi i u v} d\tilde{D}(u) , \qquad v = 0, \pm 1, \ldots$$

Usually $\tilde{d}(u)$ is a very wiggly curve; one then seeks a smooth curve $\hat{d}(t)$ which is a good (or "best") estimator of $d(t)$ .

To test whether $D(u) = u$ , $0 \le u \le 1$ , one could test equivalently whether $\varphi(v) = 0$ for $v \neq 0$ (for example, by plotting $|\tilde{\varphi}(v)|^2$ as a function of $v = 1,2,\ldots$ and determining if any of them are significantly different from zero) or whether $d(u) = 1$ , $0 \le u \le 1$ (for example, by determining if the divergence of the smoothed density

$$\hat{\Delta} = \int_0^1 \{\hat{d}(u) - 1\} \log \hat{d}(u) \, du$$

is significantly different from zero).

A method of estimating $\hat{d}(t)$ without making any prior assumptions about its behavior can be obtained using a time series prediction theoretic auto-regressive approach. The "<u>time series identification</u>" problem is to determine if there exists a difference equation of suitable order $m$ which the sequence $\varphi(v)$ satisfies:

$$\varphi(v) + a_1 \varphi(v-1) + \ldots + a_m \varphi(v-m) = 0 , \qquad v = 1,2,\ldots ,$$

where $a_m \neq 0$ . Then to test whether $D(u) = u$ , $0 \leq u \leq 1$ , one could test whether $m = 0$ . I call the problem of estimating $m$ the problem of order determination of an approximating autoregressive scheme. I propose a function of $m$ (called $CAT(m)$ , for criterion autoregressive transfer function) which can be computed from the data and is used to estimate the best order $m$ as follows; take $m$ to be the value $\hat{m}$ at which $CAT(m)$ achieves its minimum value. When $\hat{m} = 0$ we could accept the hypothesis that $D(u) = u$ or equivalently $d(u) = 1$ ; when $\hat{m} > 0$ the value of $\hat{m}$ is used to form the "optimum" smooth estimator $\hat{d}(u)$ of $d(t)$ .

When testing the fit of a model it seems desirable to use a test which indicates how to fix the model when it is found not to fit. To adapt an aphorism, such a model-testing procedure is said to have "the seeds of its own construction (rather than only destruction)."

## Preface

The typical problem facing the applied statistician (the applied statistics problem?) has been described [Easterling (1976)] as follows: "given some data, including information about how the data were obtained, what probability model(s), including parameter values, can be found which adequately explain, or describe, the data?"

I would call the foregoing a statistical science problem, and would describe it succintly to be: "model probabilities from data." A routine applied statistics problem could be formulated: "infer parameters of probability laws from data." Statisticians might not disagree that the aim of statistics should be to model probabilities by identifying (rather than

assuming) their probability laws but they might doubt whether such an aim can be realized in practice, especially with small samples.

The aim of this paper is to propose an approach to non-parametric statistical continuous data science which seems to be consistent with the conventional theories and methods of non-parametric inference but seems to point the way to universally applicable procedures (for continuous data) which are asymptotically as efficient as the best conventional goodness of fit and parameter estimation procedures available for each particular problem. We have programmed the methods described and found them successful in test cases. However, in the space available to this paper we are only able to discuss (without proofs or examples) "Chapter 1" of our work which outlines the "ideas" : how the basic general applied problems of statistical inference can be formulated as problems of estimation of distribution functions on the unit interval (or the unit hyper-cube), how such problems are more fruitfully treated as density estimation problems, and how to solve density estimation problems one can use the method which is the essence of the highly successful maximum likelihood method of parameter estimation: using a suitable information-theoretic divergence distance between densities, find the "smooth" density which is closest to a "raw" estimator of the density.

# Chapter 1

## DENSITY ESTIMATION FORMULATION OF BASIC STATISTICAL INFERENCE PROBLEMS

The aim of this chapter is to introduce a single canonical problem to which one can transform many basic statistical inference and statistical data analysis problems. This canonical problem is most simply described as the problem of testing for white noise via density estimation or smoothing. We first state some of the inference problems which we seek to unify.

One-sample (univariate) inference problems. Let $X_1,\ldots,X_n$ be i.i.d. (independent identically distributed) random variables with common a.c. (absolutely continuous) d.f. (distribution function) $F(x)$ and probability density function $f(x)$. One seeks to efficiently:

(i) estimate $f(x)$ non-parametrically (without making any prior assumption about its functional form)

(ii) test for a specified probability density $f_0(x)$ whether there exists constants $\mu$ and $\sigma$ such that

$$f(x) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right), \quad F(x) = F_0\left(\frac{x-\mu}{\sigma}\right) .$$

(iii) estimate the parameters $\mu$ and $\sigma$ (called location and scale parameters).

Two-sample (univariate) inference problems. Let $X_1,\ldots,X_n$ be i.i.d. with common a.c. d.f. $F(x)$ and let $Y_1,\ldots,Y_n$ be i.i.d. with common a.c. d.f. $G(x)$. One seeks to efficiently:

(i) test whether there exists constants $\mu$ and $\sigma$ such that

$$G(x) = F\left(\frac{x-\mu}{\sigma}\right) \quad ;$$

(ii) estimate $\mu$ and $\sigma$ .

One-sample multivariate inference problems. Let

$$\underline{X} = \begin{pmatrix} X_1 \\ \cdots \\ X_d \end{pmatrix}$$

be a random vector with absolutely continuous multivariate distribution function $F(x_1,\ldots,x_d)$ and density $f(x_1,\ldots,x_d)$ ; let $\underline{X}_1,\ldots,\underline{X}_n$ be a random sample. One seeks to efficiently:

(i) test whether the components $X_1,\ldots,X_d$ are independent random variables,

(ii) estimate the multivariate density $f$ ,

(iii) estimate the regression function

$$\mu(x_1,\ldots,x_{d-1}) = E[X_d | X_1 = x_1,\ldots,X_{d-1} = x_{d-1}]$$

In addition, there are multi-sample univariate inference problems and multi-sample multivariate inference problems concerned with the equality of many distributions; however, they are not discussed in this paper.

A CANONICAL PROBLEM (OF DENSITY ESTIMATION AND TESTING FOR CONSTANT DENSITY): One seeks to form, from "raw" estimators $\tilde{D}(u)$ , $\tilde{d}(u)$ , $\tilde{\varphi}(v)$ , "optimal" estimators $\hat{D}(u)$ , $\hat{d}(u)$ , $\hat{\varphi}(v)$ of unknown functions $D(u)$ , $d(u)$ ,

$\varphi(v)$ where (i) $D(u)$, $0 \le u \le 1$, is an absolutely continuous distribution function on the unit interval satisfying $D(0) = 0$, $D(1) = 1$; (ii) $d(u) = D'(u)$ is its density function satisfying $\log d(u)$ and $d^{-1}(u)$ are integrable functions on $0 \le u \le 1$; (iii) $\varphi(v)$ is the Fourier-Stieltjes transform

$$\varphi(v) = \int_0^1 e^{2\pi i u v} \, dD(u)$$

satisfying conditions such as $\sum_{v=-\infty}^{\infty} |\varphi(v)| < \infty$ and more generally for some $r > 0$ $\sum_{v=-\infty}^{\infty} |v|^{2r} |\varphi(v)|^2 < \infty$.

One often defines $D(u)$, $d(u)$, $\varphi(v)$ so that a "null" hypothesis is equivalent to "white noise" in the sense that the null hypothesis is equivalent to the following three equivalent conditions:

$$D(u) = u, \quad 0 \le u \le 1 ;$$

$$d(u) = 1, \quad 0 \le u \le 1 ;$$

$$\varphi(v) = 0 \quad \text{for} \quad v \neq 0 .$$

The raw estimators are usually obtained in practice by forming first either $\tilde{D}(u)$ or $\tilde{\varphi}(v)$. Then the other is formed to satisfy

$$\tilde{\varphi}(v) = \int_0^1 e^{2\pi i u v} \, d\tilde{D}(u)$$

A CANONICAL SOLUTION (OF DENSITY ESTIMATION AND TESTING FOR A CONSTANT DENSITY): Often from the observed data one can form a number $N$ of values

$d_j$ where $j = 0,1,\ldots,N-1$ which represent the jumps at the points $j/N$ in the unit interval $0 \le u \le 1$ of a raw distribution function $\tilde{D}(u)$, $0 \le u \le 1$. The Fourier transform $\tilde{\varphi}(v)$ can then be found by

$$\tilde{\varphi}(v) = \sum_{j=0}^{N-1} d_j \exp\left(2\pi i v \frac{j}{N}\right)$$

Based on $\tilde{\varphi}(\cdot)$ one computes a criterion (called CAT) which determines smooth estimators $\hat{d}(u)$, $\hat{D}(u)$, $\hat{\varphi}(v)$.

Conventional statistical methods test the null hypothesis $H_0 : D(u) = u$ by examining the deviations from zero of $\tilde{D}(u) - u$ or $\tilde{\varphi}(v)$. We accept $H_0$ if $\hat{d}(u) = 1$ or if $\hat{d}(t)$ is not significantly different from zero using the divergence

$$\hat{\Delta} = \int_0^1 \{\hat{d}(u) - 1\} \log \hat{d}(u)\, du \quad ;$$

otherwise $\hat{d}(t)$ provides an estimator of $d(t)$.

The aim of this paper is to show how to formulate diverse statistical questions so that their answer is provided by the foregoing "solution."

New parameter estimation criteria (which generate old familiar estimators in cases where they should) can be formulated using the above structure. In parametric inference one assumes a family of possible probability laws specified by probability density functions $f(x,\theta)$ indexed by a parameter $\theta$ ; to each $\theta$ one can determine a corresponding density $d_\theta(u)$, $0 \le u \le 1$, where the subscript $\theta$ indicates that it is a function of the parameter $\theta$.

Define the (raw) information divergence [compare Kullback (1958)]

$$J(\theta) = \int_0^1 - \log d_\theta(u)\, d\tilde{D}(u)$$

The proposed estimator of $\theta$ (called a minimum divergence estimator) is the value $\hat{\theta}$ at which $J(\theta)$ achieves its minimum value. Maximum likelihood estimators $\hat{\theta}$ of a parameter $\theta$ from a random sample can be defined as the values minimizing a criterion of similar form, namely

$$L(\theta) = \int_{-\infty}^{\infty} - \log f(x,\theta)\, dF_n(x) = -\frac{1}{n} \sum_{i=1}^{n} \log f(X_i,\theta) ,$$

where $F_n(x)$ is the empirical distribution function. It appears plausible that a theory of minimum divergence estimators can be developed which would parallel the theory of maximum likelihood estimators (including robustness considerations, which correspond to integrating $\log d_\theta(u)$ over a sub-interval $\epsilon \le u \le 1 - \epsilon$ ).

Another criterion useful for forming parametric estimators from densities defined over the unit interval is: choose $\theta$ to minimize

$$H(\theta) = \int_0^1 \log d_\theta(u)\, du + \int_0^1 \{d_\theta(u)\}^{-1}\, d\tilde{D}(u)$$

When applied to finite parametric normal stationary time series models, this criterion generates asymptotically efficient estimators.

When criteria yield equivalent results, we should suspect that they are calculating essentially the same thing; I believe one can show this to be the

case here in the sense that log likelihood of the "sufficient statistics" is asymptotically (and up to a constant multiplier) equal to $J(\theta)$ in the two sample and multivariate cases, and equal to $H(\theta)$ in the one univariate sample and univariate time series cases.

The non-parametric estimators $\hat{d}(u)$ which we propose for $d(u)$ are called autoregressive estimators; they are approximators to $d(u)$ expressed in terms of a parametric family of densities $d_\theta(u)$ of the form

$$d_\theta(u) = \sigma_m^2 \left| 1 + \alpha_1 e^{2\pi i u} + .. + \alpha_m e^{2\pi i u m} \right|^{-2}$$

for parameters $m, \sigma_m^2, \alpha_1, \ldots, \alpha_m$ to be estimated. Autoregressive estimators are easily evaluated at all $u$ in $0 \le u \le 1$, and easily provide estimators of derivatives and integrals of the density $d(u)$.

## 1. One Sample Statistical Inference

To identify the continuous distribution function $F(x)$ of a random sample $X_1, \ldots, X_n$ one should form first the EDF (empirical distribution function)

$$F_n(x) = \frac{1}{n} \sum_{j=1}^{n} e(x - X_j), \quad -\infty < x < \infty,$$

where

$$e(x) = 1 \quad \text{if} \quad x \geq 0$$

$$= 0 \quad \text{if} \quad x < 0 \quad .$$

In other words, $F_n(x)$ is the fraction of observations less than or equal to $x$. The inverse distribution function $F^{-1}(u)$ (also called the quantile function, in which case it is denoted $Q(u)$) of $F(x)$ is defined by

$$Q(u) = F^{-1}(u) = \inf \{x : F(x) \geq u\} , \quad 0 \leq u \leq 1$$

The quantile function has the basic property $FQ(u) = u$. The EQF (empirical quantile function) is defined by

$$Q_n(u) = F_n^{-1}(u) = \inf \{x : F_n(x) \geq u\} , \quad 0 \leq u \leq 1 \quad .$$

We show that it provides a powerful approach to test the hypothesis $H_0 : F(x) = F_0\left(\frac{x - \mu}{\sigma}\right)$ for some real $\mu$ and $\sigma > 0$ where $F_0(\cdot)$ is a specified distribution function and $\mu$ and $\sigma$ are unknown parameters (ultimately to be estimated). In terms of quantile functions one can express $H_0$ as follows:

$$H_0 : Q(u) = \mu + \sigma Q_0(u) \quad \text{for some real } \mu \text{ and } \sigma > 0$$

To prove this formula for $Q(u)$, write $x = Q(u)$ iff $F(x) = u$ iff $F_0\left(\frac{x - \mu}{\sigma}\right) = u$ iff $\frac{x - \mu}{\sigma} = Q_0(u)$ .

The existence of the derivative $f(x)$ of $F(x)$ implies the existence of the derivative, denoted $q(u)$, of $Q(u)$. Further $F\Big(Q(u)\Big) = u$ implies

$$f\Big(Q(u)\Big)\, q(u) = 1$$

We call $q(u)$ the quantile-density function and introduce the density-quantile function

$$fQ(u) = f\Big(Q(u)\Big) \quad .$$

For any $p$ in $0 < p < 1$ and $u$ in $0 < u < 1$

$$Q(u) - Q(p) = \int_p^u q(s)\, ds \quad .$$

Therefore the hypothesis $H_0$ is equivalent to the hypothesis $H_0'$ defined in terms of quantile-density functions or density-quantile functions:

$$H_0' : q(u) = \sigma\, q_0(u) \qquad \text{for some } \sigma > 0$$

or

$$H_0' : fQ(u) = \frac{1}{\sigma}\, f_0 Q_0(u) \qquad \text{for some } \sigma > 0$$

The concepts are now all assembled to show how to formulate the classic goodness of fit problem (testing $H_0$) as a density estimation problem.

Define

$$\bar{D}(u) = \int_0^u f_0 Q_0(s)\, dQ(s) = \int_0^u \frac{f_0 Q_0(s)}{fQ(s)}\, ds$$

$$D(u) = \frac{1}{\sigma_0} \int_0^u f_0 Q_0(s)\, dQ(s) = \frac{1}{\sigma_0} \int_0^u \frac{f_0 Q_0(s)}{fQ(s)}\, ds$$

defining

$$\sigma_0 = \int_0^1 \frac{f_0 Q_0(s)}{fQ(s)}\, ds$$

The null hypothesis $H_0'$ is then equivalent to $\sigma_0 = \sigma$, $\bar{D}(u) = \sigma u$ and $D(u) = u$.

Natural "raw" estimators are

$$\tilde{\bar{D}}(t) = \int_0^t f_0 Q_0(s)\, dQ_n(s) \quad ,$$

$$\tilde{D}(t) = \tilde{\sigma}_0^{-1} \int_0^t f_0 Q_0(s)\, dQ_n(s) \quad ,$$

defining

$$\tilde{\sigma}_0 = \int_0^1 f_0 Q_0(s)\, dQ_n(s) \quad .$$

These formulas are easily computed in terms of the <u>order statistics</u>

$$X_{(1)} < X_{(2)} < \ldots < X_{(n)} \quad ,$$

which are the values $X_1,\ldots,X_n$ rearranged in increasing size, since explicitly the EQF $Q_n(u)$ is given by

$$Q_n(u) = X_{(0)} \qquad \text{for } u = 0$$

$$= X_{(j)} \qquad \text{for } \frac{j-1}{n} < u \le \frac{j}{n} \quad (j = 1,\ldots,n)$$

$$= X_{(n+1)} \qquad \text{for } 1 < u \le \frac{n+1}{n} \ .$$

where $X_{(0)} = XL$ and $X_{(n+1)} = XU$ are values (which could be $-\infty$ or $\infty$) representing our prior judgment of the lower bound and upper bound of the probability distribution. Note $Q_n(u)$ is a piecewise constant function with jumps at $j/n$, $j = 0,1,\ldots,n$, of size $X_{(j+1)} - X_{(j)}$.

Spacings. If $X_1,\ldots,X_n$ is a random sample of a continuous random variable $X$, with order statistics denoted $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$, its spacings are defined by [compare Pyke (1972)]

$$q_{j,n} = n(X_{(j+1)} - X_{(j)}) \ , \qquad j = 0,1,\ldots,n-1$$

where $X_{(0)}$ is a suitable chosen finite number, and its modified spacings are defined by

$$d_{j,n} = f_0Q_0\left(\frac{j}{n}\right) q_{j,n} \ , \qquad j = 0,1,\ldots,n-1 \ ,$$

where $f_0Q_0(u)$ is a specified density-quantile function.

Non-parametric raw estimators of the distribution function $F(x)$ and quantile function $Q(u)$ are $F_n(x)$ and $Q_n(u)$ respectively. The spacing $q_{j,n}$ is a difference quotient of $Q(u)$ at $u = j/n$ and, therefore, can be regarded as a raw estimator of $q(u)$ at $u = j/n$. However, $q_{j,n}$ is not by itself a consistent estimator of $q\left(\frac{j}{n}\right)$. Consistent (and perhaps "efficient") estimators of $q(u)$ can be obtained using time series theoretic methods. More importantly our methods of estimation of $q(u)$, and therefore $fQ(u)$, yield not only their values at individual points $u$, but also various functionals (including derivatives and integrals) which are needed for adaptive and robust statistical data analysis.

These methods extend readily to censored observations and subsets of order statistics; therefore they have applications in biometry and reliability theory.

Our approach to solving the basic statistical inference questions given a random sample $X_1,\ldots,X_n$ can now be summarized as follows:

1. To non-parametically estimate the unknown probability density function $f(x)$ first non-parametrically estimate the unknown density-quantile function $fQ(u)$ through estimating the ratio

$$\bar{d}(u) = \frac{f_0Q_0(u)}{fQ(u)}$$

where $f_0Q_0(u)$ is a specified density-quantile chosen to "guarantee" that $\bar{d}(u)$ have various integrability properties whose necessity will arise in the course of our theoretical development.

2. To test whether a specified $f_0(\cdot)$ is the true probability density (up to location and scale parameters $\mu$ and $\sigma$) choose the corresponding density-quantile function $f_0Q_0(u)$ as the function to be used in forming from spacings raw estimators $\tilde{D}(u)$ and $\tilde{\varphi}(v)$ and tests of the hypothesis that the density $\bar{d}(u)$ is a constant function.

3. To form efficient estimators $\hat{\mu}$ and $\hat{\sigma}$ of the location and scale parameters it suffices to know (or to have estimated) $f_0Q_0(u)$ and $Q_0(u)$ since then one treats the estimation as a problem of regression on a continuous parameter time series using the fact that, as $n \to \infty$, the asymptotic distribution [compare Shorack (1972)] of

$$\sqrt{n}\ fQ(u)\{Q_n(u) - Q(u)\} = \sqrt{n}\ \frac{1}{\sigma}\ f_0Q_0(u)\{Q_n(u) - \mu - \sigma Q_0(u)\}$$

is the Brownian bridge $B(u)$ which is a normal zero mean stochastic process with covariance kernel $E[B(s)\ B(t)] = \min(s,t) - st$. Estimators $\hat{\mu}$ and $\hat{\sigma}$ are then of the usual regression analysis form [compare Parzen (1961), (1970)]

$$\begin{pmatrix} \hat{\mu} \\ \\ \hat{\sigma} \end{pmatrix} = \mathrm{Inf}_0^{-1} \begin{pmatrix} T_{n,\mu} \\ \\ T_{n,\sigma} \end{pmatrix} .$$

The information matrix $\mathrm{Inf}_0$ is defined by

$$\mathrm{Inf}_0 = \begin{bmatrix} < f_0Q_0\ ,\ f_0Q_0 > & < f_0Q_0\ ,\ Q_0(f_0Q_0) > \\ \\ < Q_0(f_0Q_0)\ ,\ f_0Q_0 > & < Q_0(f_0Q_0)\ ,\ Q_0(f_0Q_0) > \end{bmatrix}$$

in terms of a (reproducing kernel Hilbert space) inner product

$$< f,g > = \int_0^1 f'(t)\, g'(t)\, dt = -\int_0^1 f''(t)\, g(t)\, dt$$

between differentiable functions $f(t)$ and $f(t)$ satisfying $f'(0)\, g(0) = f'(1)\, g(1) = 0$ . The statistics $T$ are linear combinations of order statistics found as follows:

$$T_{n,\mu} = < f_0Q_0 \,,\, Q_n(f_0Q_0) >$$

$$= -\int_0^1 Q_n(t)\; f_0Q_0(t)\; (f_0Q_0)''(t)\; dt \quad .$$

$$T_{n,\sigma} = < Q_0(f_0Q_0) \,,\, Q_n(f_0Q_0) >$$

$$= -\int_0^1 Q_n(t)\; f_0Q_0(t) \{Q_0(f_0Q_0)\}''(t)\; dt$$

Explicitly,

$$T_{n,\mu} = \sum_{j=1}^{n} X_{(j)} \{W_\mu\left(\frac{j}{n}\right) - W_\mu\left(\frac{j-1}{n}\right)\} \quad ,$$

$$W_\mu(u) = -\int_0^u f_0Q_0(s)(f_0Q_0)''(s)\; ds \quad ,$$

$$W_\mu'(u) = f_0Q_0(u)\; J_0'(u) \quad ;$$

$$T_{n,\sigma} = \sum_{j=1}^{n} X_{(j)} \{W_\sigma\left(\frac{j}{n}\right) - W_\sigma\left(\frac{j-1}{n}\right)\} \quad ,$$

$$W_\sigma(u) = -\int_0^u f_0 Q_0(s)\{Q_0(f_0 Q_0)\}''(s)\, ds \quad ,$$

$$W_\sigma'(u) = J_0(u) + Q_0(u)\, W_\mu'(u) \quad .$$

The function $J_0(u)$ is defined by

$$J_0(u) = -\frac{d}{du} f_0 Q_0(u)$$

and is called the score function. It plays a basic role in the theory of nonparametric estimation, and is most easily estimated using the fact that it is the derivative of the density-quantile function, rather than the formula

$$J_0(u) = -\frac{f_0'\left(Q_0(u)\right)}{f_0\left(Q_0(u)\right)} = -\frac{f_0'\left(F_0^{-1}(u)\right)}{f_0\left(F_0^{-1}(u)\right)}$$

A list of density-quantile functions and score functions of familiar univariate continuous probability laws is given in Table I.

## 2. Tests for the Equality of Two Distributions

This section introduces a density estimation approach to non-parametric tests of the hypothesis $H_0$ that two independent samples (a random sample $X_1,\ldots,X_m$ of a continuous random variable $X$, and a random sample $Y_1,\ldots,Y_n$ of a continuous random variable $Y$) are drawn from identical populations in the sense that $X$ and $Y$ are identically distributed; in symbols,

$$H_0 : F_X(x) = F_Y(x) \quad \text{for all} \quad x \quad \text{in} \quad -\infty < x < \infty \quad .$$

One way to define a distribution function $D(u)$, $0 \le u \le 1$ such that $H_0$ is equivalent to the hypothesis $H_0' : D(u) = u$, $0 \le u \le 1$, is to define

$$D(u) = F_X\left(Q_Y(u)\right) \qquad \text{or} \qquad D(u) = F_Y\left(Q_X(u)\right)$$

Such statistics remain to be investigated. A statistic which corresponds to currently used tests of $H_0$ is obtained by defining

$$H(x) = \lambda F_X(x) + (1 - \lambda) F_Y(x)$$

where $\lambda$ is the limit of $\frac{m}{m+n}$, the fraction of $X$ values in the combined samples of $X$ and $Y$ values. In words, $H(x)$ is a mixture of the distributions of $X$ and $Y$.

Denote by $F_{X,m}(x)$ and $F_{Y,n}(x)$ the EDF of the $X$ and $Y$ samples, and let $H_N(x)$ be the EDF of the combined samples of $X$ and $Y$ values,

where $N = m + n$ . Then, defining $\lambda_N = \frac{m}{N}$ ,

$$H_N(x) = \lambda_N F_{X,m}(x) + (1 - \lambda_N) F_{Y,n}(x) \quad .$$

Since $F_X(x) = F_Y(x) = H(x)$ under $H_0$ , one can test this hypothesis by testing the uniformity of

$$D(u) = F_X\left(H^{-1}(u)\right) \quad , \qquad 0 \le u \le 1$$

whose natural raw estimator is

$$\tilde{D}(u) = F_{X,m}\left(H_N^{-1}(u)\right)$$

This approach can be readily extended to testing the equality of $k$ samples of random variables $X_1,\ldots,X_k$ , if one considers for $j = 1,\ldots,k$ ,

$$D_j(u) = F_{X_j}\left(H^{-1}(u)\right)$$

where $H(x)$ is the distribution function of the combined sample.

Now $H_N^{-1}(u)$ is a piecewise constant distribution function whose value in the interval $\left(\frac{k-1}{N}, \frac{k}{N}\right]$ is the k-th value in the combined sample. Therefore, for $j = 1,\ldots,m-1$

$$\tilde{D}(t) = \frac{j}{m} \quad \text{for} \quad \frac{R(X_{(j)}) - 1}{N} < u \le \frac{R(X_{(j+1)}) - 1}{N}$$

$$= 1 \quad \text{for} \quad \frac{R(X_{(m)}) - 1}{N} \le u < 1$$

where $R(X_i)$ is the rank of $X_i$ as a member of the combined sample $X_1,\ldots,X_m, Y_1,\ldots,Y_n$. In words, $\tilde{D}(u)$ is a piecewise constant distribution function with jumps of size $1/m$ at all points $u$ of the form $u = \{R(X_{(j)}) - 1\}/N$, $j = 1,2,\ldots,m$.

The Fourier transforms

$$\varphi(v) = \int_0^1 e^{2\pi iuv}\, dD(u)$$

have natural raw estimators

$$\tilde{\varphi}(v) = \int_0^1 e^{2\pi iuv}\, d\tilde{D}(u)$$

$$= \frac{1}{m} \sum_{j=1}^{m} \exp\left\{2\pi iv\, \frac{R(X_{(j)}) - 1}{N}\right\}$$

Many statistics (denoted $T_N$, where $N = m + n$ is the total sample size) which have been suggested to test $H_0$ are linear combinations of the rank-order statistics $R(X_i)$. Chernoff and Savage introduced a representation for linear rank statistics in a pioneering paper (1958):

$$T_N = m \int_{-\infty}^{\infty} J_N[H_N(x)]\, dF_{X,m}(x)$$

$$= \sum_{i=1}^{m} J_N\left(\frac{R(X_i)}{N}\right)$$

where $J_N(t)$ is a score function which tends, as $N \to \infty$, "suitably" to a limit $J(t)$. The foregoing representation of $T_N$ may be written (by

suitably defining $J_N$)

$$T_N = m \int_0^1 J_N(u)\, d\tilde{D}(u) \quad .$$

Let us show how test statistics of this form arise from our point of view.

To test the hypothesis

$$H_0 : d(u) = 1$$

against a simple alternative

$$H_1 : d(u) = d_1(u)$$

where $d_1(u)$ is a specified function one can show that an asymptotic likelihood ratio test statistic is the "correlator"

$$R_1 = \int_0^1 \{d_1(t) - 1\}\, d\tilde{D}(t)$$

Now suppose that the alternative family of densities is denoted $d_\theta(t)$ to indicate that it is parametrized by a parameter $\theta$ ; suppose we have the expansion

$$d_\theta(t) = 1 + \theta\,\delta(t) \quad \text{for} \quad \theta \quad \text{near zero}$$

where

$$\delta(t) = \frac{\partial}{\partial\theta}\, d_\theta(t)\bigg|_{\theta = 0} \quad .$$

The "correlator" statistic $R_1$ as a likelihood ratio statistic for testing $H_1$ against $H_0$ is then equivalent to the linear detector

$$R_\delta = \int_0^1 \delta(t)\, d\tilde{D}(t)$$

for $\theta$ near zero. By a direct calculation of $\delta(u)$ below we show that to test $H_0 : F_X(x) = F_Y(x) = F(x)$ against the alternative

$$H_1 : F_X(x) = F(x),\ F_Y(x) = F(x - \theta)$$

the best test for $\theta$ close to 0 is based on the statistic $R_\delta$ where

$$\delta(u) = -(1 - \lambda)\, J(u)$$

where $J(u)$ is the score function

$$J(u) = -\frac{d}{du} fQ(u)$$

Assume that the density $f(x)$ is a symmetric function of $x$ ; then

$$E_\theta[R_\delta] = \int_0^1 \delta(u)\, d(u)\, du = \theta \int_0^1 \delta^2(u)\, du$$

$$= \theta \int_0^1 (1 - \lambda)^2\, J^2(u)\, du$$

so that an approximately unbiased estimator of $\theta$ is

$$\theta^* = \int_0^1 J(u)\, d\tilde{D}(u) \div (1 - \lambda) \int_0^1 J^2(u)\, du$$

whose variance may be shown to be approximately equal to

$$\text{Var } (\theta^*) = \frac{1}{N}\{\lambda(1 - \lambda) \int_0^1 J^2(u)\, du\}^{-1}$$

which is also the variance of the maximum likelihood estimator.

When the linear detector $R_\delta$ is used to test whether the location parameter $\theta$ equals 0, one accepts this hypothesis (for large sample sizes N) if

$$\frac{|\theta^*|^2}{N \text{ Var } (\theta^*)} = \frac{R_\delta^2}{N \text{ Var } (R_\delta^2)} = \frac{\lambda\{\int_0^1 \delta(t)\, d\tilde{D}(t)\}^2}{(1 - \lambda)\int_0^1 \delta^2(t)\, dt}$$

is below a suitable threshold (one can argue that the threshold is a number of the form C/N where C is often 2 or 4). I am proposing that instead of $R_\delta$ one use a non-parametric estimator $\hat{\Delta}$ of the divergence

$$\Delta = \int_0^1 \{d_\theta(u) - 1\} \log d_\theta(u)\, du = \theta^2 \int_0^1 \delta^2(t)\, dt \quad ;$$

if $\theta$ were estimated by $\theta^*$, let $\Delta$ be denoted by $\Delta^*$:

$$\Delta^* = \left|\int_0^1 \delta(t)\, d\tilde{D}(t)\right|^2 \div \int_0^1 \delta^2(t)\, dt$$

It seems plausible that the proposed "universal" test of the hypothesis $\theta = 0$, which accepts it when $\hat{\Delta}$ is below a suitable threshold of the form of C/N

- (for a suitable value of the constant C), would perform as well as the best test of the form of $R_\delta$ since it appears to be asymptotically calculating (up to a constant multiplier) the same statistic!

To calculate $\delta(u)$ we must calculate the density $d_\theta(u)$ corresponding to the canonical distribution function

$$D_\theta(u) = FH_\theta^{-1}(u)$$

where

$$H_\theta(x) = \lambda F_X(x) + (1 - \lambda) F_Y(x) = \lambda F(x) + (1 - \lambda) F(x - \theta) \quad .$$

To establish a formula for $d_\theta(u)$ we obtain from the defining equation for $D_\theta(u)$ that

$$H_\theta F^{-1} D_\theta(u) = u \quad ,$$

whence

$$u = \lambda D_\theta(u) + (1 - \lambda) F(QD_\theta(u) - \theta)$$

where $Q(u) = F^{-1}(u)$ is the quantile function. Differentiating with respect to u

$$1 = \lambda d_\theta(u) + (1 - \lambda)\, f(QD_\theta(t) - \theta)\, q\Big(D_\theta(u)\Big)\, d_\theta(u)$$

whence

$$\{d_\theta(u)\}^{-1} = \lambda + (1 - \lambda)\, f(QD_\theta(u) - \theta)\, q\Big(D_\theta(u)\Big)$$

Differentiating with respect to $\theta$ :

$$-\{d_\theta(u)\}^{-2} \frac{\partial}{\partial\theta} d_\theta(u) = (1 - \lambda)\, f(QD_\theta(u) - \theta)\, q'\Big(D_\theta(u)\Big) \frac{\partial}{\partial\theta} D_\theta(u)$$

$$+ (1 - \lambda)\, q\Big(D_\theta(u)\Big)\, f'(QD_\theta(u) - \theta)\{q\Big(D_\theta(u)\Big) \frac{\partial}{\partial\theta} D_\theta(u) - 1\}$$

Setting $\theta = 0$ and using the identities $q(u)\, f'\Big(Q(u)\Big) = (fQ)'(u)$, $fQ(u)\, q'(u) + (fQ)'(u)\, q(u) = 0$ , one obtains the desired conclusion: $\delta(u) = -(1 - \lambda)\, J(u)$ .

To test a scale parameter $\theta$ , one considers alternative hypotheses $G(x) = F(x\theta)$ , where $\theta = 1$ represents the null hypothesis. Using the foregoing argument one can show that

$$\delta(u) = -(1 - \lambda)\{Q(u)\, fQ(u)\}'$$

$$= (1 - \lambda)\{Q(u)\, J(u) - 1\}$$

Asymptotic variance of linear rank statistics. In terms of the canonical distribution function $D(u)$ and its density $d(u)$ , we can obtain rather simple formulas for the asymptotic variance $\sigma^2$ of the linear rank statistics of the form

$$T_N = \int_0^1 J(u)\, d\tilde{D}(u) \quad ,$$

which satisfy the conditions of the Chernoff-Savage theorem (1958);

$\sqrt{N}\ (T_N - \mu)$ is asymptotically $N(0,\sigma^2)$ where $\mu = \int_0^1 J(t)\ dD(t)$

$$\sigma^2 = 2 \int_0^1 \int_0^1 ds\ dt\ e(t-s)\ J'(s)\ J'(t)$$

$$\{\frac{1}{1-\lambda}\ \Big(s - \lambda D(s)\Big)\Big(1 - t - \lambda\big(1 - D(t)\big)\Big)\ d(s)\ d(t)$$

$$+ \frac{1}{\lambda}\ D(s)\big(1 - D(t)\big)\big(1 - \lambda d(s)\big)\big(1 - \lambda d(t)\big)\}$$

Under the null hypothesis $D(u) = u$

$$\mu = \int_0^1 J(t)\ dt$$

$$\sigma^2 = 2\ \frac{1-\lambda}{\lambda} \int_0^1 \int_0^1 ds\ dt\ e(t-s)\ J'(s)\ J'(t)\ s(1-t)$$

$$= \Big(\frac{1-\lambda}{\lambda}\Big) \int_0^1 \{J(t) - \int_0^1 J(s)\ ds\}^2\ dt$$

An important extension of these results is to $J(u) = e^{2\pi i v u}$ ; one obtains that under the null hypothesis of independence, $\{\tilde{\varphi}(v)\ ,\ v = \pm 1, \pm 2, \ldots\}$ are asymptotically independent $N\Big(0, \frac{1}{N}\Big)$ .

## 3. Tests for Independence, Multivariate Density Estimation, and Non-Parametric Regression

When the data consist of a random sample $\underline{X}_1,\ldots,\underline{X}_n$ of an m-dimensional random vector

$$\underline{X} = \begin{bmatrix} X^1 \\ \vdots \\ X^m \end{bmatrix}$$

it is often of interest to test the hypothesis $H_0 : X^1,\ldots,X^m$ are independent random variables.

Let the joint distribution function and probability density of $\underline{X}$ be denoted by $F(x_1,\ldots,x_m)$ and $f(x_1,\ldots,x_m)$ respectively. Let its marginal distribution functions and densities be denoted by $F_k(x_k)$ and $f_k(x_k)$. Note that $f_k(x_k)$ is the probability density of the k-th component $X^k$. Corresponding to each density $f_k(x_k)$ there is a quantile function $Q_k(u_k)$, and a density quantile function $f_kQ_k(u_k)$.

The hypothesis $H_0$ that the components of $\underline{X}$ are independent can be expressed

$$H_0 : F(x_1,\ldots,x_m) = F_1(x_1)\ldots F_m(x_m)$$

Equivalently,

$$H_0 : F(Q_1(u_1),..,Q_m(u_m)) = u_1\ldots u_m$$

Define

$$D(u_1,..,u_m) = F\left(Q_1(u_1),\ldots,Q_m(u_m)\right) \quad ;$$

it is the joint distribution function of the uniformly distributed random variables

$$U^1 = F_1(X^1),..,U^m = F_m(X^m)$$

We shall test $H_0$ by estimating the joint density function

$$d(u_1,\ldots,u_m) = \frac{\partial^m}{\partial u_1 \ldots \partial u_m} D(u_1,..,u_m)$$

$$= \frac{f\left(Q_1(u_1),\ldots,Q_m(u_m)\right)}{f_1Q_1(u_1)\ldots f_mQ_m(u_m)}$$

Note that in the case of a multivariate location and scale parameter family of probability densities

$$f(x_1,\ldots,x_m) = \frac{1}{\sigma_1 \cdots \sigma_m} f^0\left(\frac{x_1 - \mu_1}{\sigma_1}, \ldots, \frac{x_m - \mu_m}{\sigma_m}\right)$$

each marginal density is of the form

$$f_k(x_k) = \frac{1}{\sigma_k} f_k^0\left(\frac{x_k - u_k}{\sigma_k}\right)$$

and the individual quantile functions are of the form

$$Q_k(u_k) = \mu_k + \sigma_k Q_k^0(u_k) \quad .$$

Therefore

$$d(u_1,\ldots,u_m) = \frac{f^0\left(Q_1^0(u_1),\ldots,Q_m^0(u_m)\right)}{f_1^0 Q_1^0(t_1) \ldots f_m^0 Q_m^0(u_m)}$$

Therefore the density $d(u_1,\ldots,u_m)$ does not depend on location and scale parameters and is a measure of association or dependence. In particular $d(u_1,..,u_m)$ is identically equal to 1 if and only if all components of $\underline{X}$ are independent. An overall measure of association can be defined by the divergence

$$\Delta = \int_0^1 \ldots \int_0^1 \{d(u_1,\ldots,u_m) - 1\} \log d(u_1,..,u_m)\, du_1 \ldots du_m$$

We call: $d(u_1,..,u_m)$ the <u>regression-density</u> of $\underline{X}$ ; $\Delta$ the <u>regression-density-divergence</u>; $D(u_1,..,u_m)$ the <u>regression-distribution</u> function.

For the bivariate normal distribution with correlation coefficient $\rho$ , the regression density is given by

$$d(s,t;\rho) = (1-\rho^2)^{-\frac{1}{2}} \exp\left[\{-2(1-\rho^2)\}^{-1}\{\left|\rho\Phi^{-1}(s)\right|^2 + \left|\rho\Phi^{-1}(t)\right|^2 - 2\rho\Phi^{-1}(s)\,\Phi^{-1}(t)\}\right]$$

and the regression-density-divergence is given by

$$\Delta = \frac{\rho^2}{1-\rho^2}$$

For a multivariate normal distribution with correlation matrix $R$ (that is $R$ is the matrix of correlation coefficients between the components of the random vector) the regression-density-divergence is given by

$$\Delta = \frac{1}{2} \operatorname{tr}(R^{-1} - I)$$

To estimate regression-distribution $D$ from a sample of size $n$ we use the natural raw estimator

$$\tilde{D}(u_1,..,u_m) = F_n\left(Q_{1,n}(u_1),\ldots,Q_{m,n}(u_m)\right)$$

where $F_n(x_1,\ldots,x_m)$ is the empirical distribution function of the $n$ random vectors $\underline{X}_1,\ldots,\underline{X}_n$, and $Q_{k,n}(u_k)$ is the sample quantile function of the k-th components of these vectors.

Define the Fourier-Stieltjes transforms

$$\tilde{\varphi}(v_1\ldots v_m) = \int_0^1 \cdots \int_0^1 e^{2\pi i(v_1u_1+\ldots+v_mu_m)} \, d\tilde{D}(u_1,\ldots,u_m)$$

$$\varphi(v_1,\ldots,v_m) = \int_0^1 \cdots \int_0^1 e^{2\pi i(v_1u_1+\ldots+v_mu_m)} \, dD(u_1,\ldots,u_m)$$

$$= \int_0^1 \cdots \int_0^1 e^{2\pi i(v_1u_1+\ldots+v_mu_m)} \, d(u_1,\ldots,u_m)\, du_1\ldots du_m \quad .$$

It will be shown in the sequel that $\tilde{D}$ tends to $D$ and $\tilde{\varphi}$ tends to $\varphi$ (as $n \to \infty$) therefore we can form (using time series theoretic statistical

methods) estimators $\hat{d}$ that tend to $d$, which can be used to test whether $d$ is identically 1 (which is equivalent to the null hypothesis of independence). The multivariate case discussed in this section seems to me to demonstrate the power of reducing statistical problems to estimation of densities. If the only statistic one works with is the distribution function $D_n(u_1,\ldots,u_m)$ one is confronted with the difficult task of testing whether it is significantly different from the uniform distribution function $u_1u_2\ldots u_m$. Then if one flunks this test, and rejects the assumption of independent components of the random vector $\underline{X}$, one has no means of modeling the dependence.

The empirical regression-distribution function $\tilde{D}(u_1,\ldots,u_m)$ is a purely discrete distribution which assigns mass $1/n$ to the $n$ points $\left(\frac{R_1(X_j^1) - 1}{n},\ldots,\frac{R_m(X_j^m) - 1}{n}\right)$ which are the rank vectors of the $n$ random vectors $(X_j^1,\ldots,X_j^m)$ for $j = 1,\ldots,n$; here $R_k(X_j^k)$ denotes the rank of $X_j^k$ among $X_1^k,\ldots,X_n^k$.

Asymptotically our conclusions are unchanged if we take as our raw estimator $\tilde{D}(u_1,\ldots,u_m)$ of $D(u_1,\ldots,u_m)$ the purely discrete distribution function which assigns mass $1/n$ to the $n$ points $\left(\frac{R_1(X_j^1)}{n},\ldots,\frac{R_m(X_j^m)}{n}\right)$; then the raw estimator of $\varphi(v_1,\ldots,v_m)$ is

$$\tilde{\varphi}(v_1,\ldots,v_m) = \frac{1}{n}\sum_{j=1}^{n} \exp\left\{2\pi i\left(v_1\frac{R_1(X_j^1)}{n} + \ldots + v_m\frac{R_m(X_j^m)}{n}\right)\right\}$$

In the two-dimensional case $(m = 2)$ we denote the observed data by $(X_1,Y_1),\ldots,(X_n,Y_n)$. Then the $n$ jump points of $\tilde{D}$ are of the form $\left(\frac{j}{n},\frac{R_j}{n}\right)$ where $R_j$ is the rank among the $Y$'s of that $Y$-value corresponding

to the X-value with rank $j$. The well-known rank tests for independence (see Hajek (1969)) may be expressed in terms of the vector $R_1,\ldots,R_n$ as follows:

Spearman test $$S = \sum_{j=1}^{n} jR_j$$

Quadrant test $$S = \sum_{j\geq \frac{1}{2}n+1}^{n} e(R_j - \tfrac{1}{2}n - 1)$$

Kendall rank correlation coefficient $$K = \sum_{i=1}^{n} \sum_{j>i} e(R_j - R_i) \quad .$$

Therefore one may readily establish the connection between our time series theoretic approach to tests for independence and conventional tests.

To test the hypothesis of independence (regression density identically 1) one may be willing to assume a family of alternative hypotheses indexed by a parameter $\theta$ under which the regression-density may be represented

$$d_\theta(u_1,\ldots,u_m) = 1 + \theta\,\delta(u_1,\ldots,u_m)$$

for $\theta$ close to 0. Then an asymptotic likelihood ratio statistic for testing independence is

$$R_\delta = \int_0^1 \cdots \int_0^1 \delta(u_1,\ldots,u_m)\, d\tilde{D}(u_1,..,u_m) \quad .$$

In the case $m = 2$

$$R_\delta = \frac{1}{n} \sum_{j=1}^{n} \delta\left(\frac{j}{n}, \frac{R_j}{n}\right)$$

The regression-density function of the bivariate normal distribution, denoted $d_\rho(u_1,u_2)$, is a function of the correlation coefficient $\rho$ such that

$$\delta(u_1,u_2) = \frac{\partial}{\partial\rho} d_\rho(u_1,u_2)\Big|_{\rho=0} = \Phi^{-1}(u_1)\,\Phi^{-1}(u_2)$$

Therefore our approach yields

$$R_\delta = \frac{1}{n} \sum_{j=1}^{n} \Phi^{-1}\left(\frac{j}{n}\right)\,\Phi^{-1}\left(\frac{R_j}{n}\right)$$

as an optimum non-parametric statistic for testing independence against bivariate normal dependence. The statistic $R_\delta$ is the Fisher-Yates or normal scores statistic well studied in the theory of non-parametric statistics. The Spearman and quadrant tests are linear rank statistics corresponding to the weight functions

Spearman $\quad \delta(u_1,u_2) = u_1u_2$

Quadrant $\quad \delta(u_1,u_2) = e\left(u_1 - \frac{1}{2} - \frac{1}{n}\right)\, e\left(u_2 - \frac{1}{2} - \frac{1}{n}\right)$ .

$K$ is a linear function of $\int_0^1\int_0^1 \tilde{D}(u,v)\, d\tilde{D}(u,v)$ .

The concept of minimum divergence estimation (defined in the introduction to this chapter) can be illustrated in the present context. To estimate the correlation coefficient $\rho$ of the bivariate normal distribution, the

minimum divergence estimator $\hat{\rho}$ is the value $\rho$ at which

$$J(\rho) = \int_0^1 \int_0^1 - \log d_\rho(u_1,u_2)\, d\tilde{D}(u_1,u_2)$$

achieves its minimum. By solving $J'(\rho) = 0$ one may show that $\hat{\rho} = R_\delta$ .

Kimeldorf and Sampson (1975) list parametric bivariate regression-densities corresponding to various multivariate distributions; one could estimate their parameters using J and H divergence functionals of $\tilde{D}(u_1,u_2)$ .

Multivariate Density Estimation: An estimator $\hat{d}(u_1,u_2)$ leads to an estimator of $f(x_1,x_2)$ , using the relation

$$\hat{f}\Big(Q_1(u_1),\ Q_2(u_2)\Big) = \hat{f}_1 Q_1(u_1)\ \hat{f}_2 Q_2(u_2)\ \hat{d}(u_1,u_2) \quad .$$

Nonparametric Regression: An outstanding problem of statistics is the estimation of the non-parametric regression of $X_2$ on $X_1$ in the sense of the conditional mean

$$E[X_2|X_1 = x_1] = \int_{-\infty}^{\infty} x_2 f_{X_2|X_1}(x_2|x_1)\, dx_2$$

$$= \int_{-\infty}^{\infty} x_2 \frac{f(x_1,x_2)}{f(x_1)}\, dx_2 \quad .$$

By making the change of variable $x_2 = Q_2(u_2)$ or $u_2 = F_2(x_2)$ , we obtain

$$E[X_2|X_1 = x_1] = \int_0^1 Q_2(u_2) \frac{f(x_1, Q_2(u_2))}{f(x_1)} q_2(u_2)\, du_2 \quad ;$$

this formula can be rewritten to yield the following remarkable theorem.

Theorem: (Regression-Density Formula for Conditional Expectation and Non-Parametric Regression)

$$E[X_2|X_1 = Q(u_1)] = \int_0^1 Q_2(u_2)\, d(u_1,u_2)\, du_2$$

which justifies calling $d(u_1,u_2)$ the regression density; note that

$$d(u_1,u_2) = \frac{f\left(Q_1(u_1), Q_2(u_2)\right)}{f_1 Q_1(u_1)\, f_2 Q_2(u_2)} \quad .$$

If one estimates $Q_2(u_2)$ by the empirical quantile function $Q_{2,n}(u_2)$ the corresponding estimated conditional expectation is

$$\hat{E}[X_2|X_1 = Q(u_1)] = \sum_{j=1}^{n} X_{2,j} \int_{(j-1)/n}^{j/n} \hat{d}(u_1,u_2)\, du_2$$

$$= \sum_{j=1}^{n} X_{2,j} \left\{ \hat{D}_1\left(u_1, \frac{j}{n}\right) - D_1\left(\hat{u}_1, \frac{j-1}{n}\right) \right.$$

where $\hat{D}_1(u_1,u_2)$ is an estimator of

$$D_1(u_1,u_2) = \int_0^{u_2} d(u_1,u_2')\, du_2' = \frac{\partial}{\partial u_1} D(u_1,u_2) \quad .$$

The approach to non-parametric multivariate density estimation and

non-parametric regression outlined above (whose theory and practice remains to be investigated) appears to show that one can estimate regressions without estimating probability laws!

One often prefers to calculate regressions as conditional quantile functions; then one can proceed as follows. An expression for the conditional distribution function of $X_2$ given $X_1$ is

$$F_{X_2|X_1}(x_2|x_1) = \int_0^{u_2} d(F_1(x_1),u_2')\, du_2'$$

where $u_2 = F_{X_2}(x_2)$ . It follows that the conditional quantile function of $X_2$ given $X_1$ is given by

$$Q_{X_2|X_1}(p|x_1) = Q_2 D_1^{-1}(F_X(x_1),p)$$

In words, the conditional quantile function equals the unconditional quantile function $Q_2(u)$ with a change of variable $u = D_1^{-1}(F_X(x_1),p)$ .

While we recommend Fourier theoretic methods of estimating $D_1(u_1,u_2)$ it should be noted that a quick and dirty estimator can be provided by a "naive k-nearest neighbor" estimator

$$D_1^*(u_1,u_2) = \frac{n}{2k}\left\{\tilde{D}(u_1 + \frac{k}{n}, u_2) - \tilde{D}(u_1 - \frac{k}{n}, u_2)\right\} .$$

To understand the dramatic nature of our approach to non-parametric regression imagine a scatter diagram of points $(X_i,Y_i)$ $i = 1,2,\ldots,n$ in the plane. One seeks to fit a smooth curve $y = g(x)$ through the points. A

typical criterion of curve fitting might be to: find $g$ to minimize

$$\frac{1}{n} \sum_{j=1}^{n} \{Y_j - g(X_j)\}^2 + \lambda \int_a^b |g^{(m)}(x)|^2 \, dx$$

where $g^{(m)}$ is the m-th derivative of $g$ assumed to exist over some specified interval $a$ to $b$. The solution is then a polynomial spline of degree $2m - 1$ [see Wahba (1976)]. Rather than choose a function $g$ by such an optimization criterion (which is inevitably ad hoc and still requires one to specify $\lambda$, $m$, $a$ and $b$) we are proposing that one adopt as one's "optimal smooth curve" a curve of the form

$$y = g(x) = \int_0^1 \hat{Q}_Y(u) \, \hat{d}(\hat{F}_X(x), u) \, du$$

where $\hat{Q}_Y(u)$ is an estimator of the quantile function of the Y-values, $\hat{F}_X(x)$ is an estimator of the distribution function of X-values, and $\hat{d}(s,t)$ is the estimated regression-density function. How does one explain to a numerical analyst what are the optimizing properties of the procedure we are proposing?

Multi-dimensional non-parametric regression: The foregoing results can be extended to multi-dimensions. We state only a formula for the conditional expectation of $X_m$ on $X_1, \ldots, X_{m-1}$ :

$$E[X_m | X_1 = Q_1(u_1), \ldots, X_{m-1} = Q_{m-1}(u_{m-1})]$$

$$= \int_0^1 Q_m(u_m) \frac{d_m(u_1, \ldots, u_m)}{d_{m-1}(u_1, \ldots, u_{m-1})} \, du_m$$

where

$$d_m(u_1,..,u_m) = \frac{f\left(Q_1(u_1),...,Q_m(u_m)\right)}{f_1Q_1(u_1) \ldots f_mQ_m(u_m)}$$

is the regression-density function of $X_1,..,X_m$ (and $d_{m-1}$ is the regression-density function of $X_1,..,X_{m-1}$).

Asymptotic distribution of statistics of the form $T_n = \int_0^1\int_0^1 J(s)\ K(t)\ \tilde{dD}(s,t)$.

The work of Ruymgaart (1974) leads us to the following roughly stated limit theorem:

$$\sqrt{n}\ (T_n - \mu) \quad \text{is asymptotically} \quad N(0,\sigma^2)$$

where

$$\mu = \int_0^1\int_0^1 J(s)\ K(t)\ dD(s,t)$$

$$\sigma^2 = \int_0^1\int_0^1 |V(s,t)|^2\ dD(s,t)$$

$$V(s,t) = J(s)\ K(t) - \int_0^1\int_0^1 J(u)\ K(v)\ dD(u,v)$$

$$+ \int_0^1\int_0^1 [e(u-s) - u]\ J'(u)\ K(v)\ dD(u,v)$$

$$+ \int_0^1\int_0^1 [e(v-t) - v]\ J(u)\ K'(v)\ dD(u,v)$$

Under the null hypothesis $D(s,t) = st$ ,

$$\mu = \int_0^1 J(s)\, ds \int_0^1 K(t)\, dt$$

$$\sigma^2 = \int_0^1 \int_0^1 |V(s,t)|^2\, ds\, dt$$

$$V(s,t) = \{J(s) - \int_0^1 J(u)\, du\}\{K(t) - \int_0^1 K(v)\, dv\}$$

Extending these results to $J(s) = e^{2\pi i v_1 s}$ , $K(t) = e^{2\pi i v_2 t}$ one obtains that under the null hypothesis of independence $\{\tilde{\varphi}(v_1,v_2)\,,\ v_1,v_2 = \pm 1, \pm 2,\ldots\}$ are asymptotically independent $N\left(0,\frac{1}{n}\right)$ .

Joint distribution of the sample quantile functions of two dependent random variables $X^1$ and $X^2$ . It has been noted in Section 1 that the modified empirical quantile function deviations

$$\bar{Q}_{j,n}(u) = \sqrt{n}\ f_j Q_j(u)\{Q_{j,n}(u) - Q_j(u)\}$$

is asymptotically $N\left(0,t(1-t)\right)$ ; further $\bar{Q}_{j,n}(s)$ and $\bar{Q}_{j,n}(t)$ have asymptotic covariance $s(1-t)$ when $s < t$ . Weiss (1964) proves that asymptotically

$$\mathrm{Cov}\left(\bar{Q}_{j,n}(s)\,,\bar{Q}_{k,n}(t)\right) = D_{jk}(s,t) - st\ ,\quad s \le t\ ,$$

defining

$$D_{jk}(s,t) = F_{jk}\Big(Q_j(s),\ Q_k(t)\Big) \quad .$$

Using this result, one could obtain asymptotically efficient unbiased estimators, from incomplete samples, of the common mean $\mu$ of bivariate normal random variables $X$ and $Y$ with unknown unequal variances and unknown covariance (for other estimators, see Hamdan, Pirie, and Khuri (1976)).

## 4. Time Series Analysis and Autoregressive Model Approximation

The density estimation problem (which we claim is a canonical problem to which one can transform many basic problems of statistical inference) first arose in the analysis of stationary time series.

Let $Y(t)$, $t = 0, \pm 1, \pm 2, \ldots$ be a zero mean covariance stationary normal time series; its probability law is then specified by either the covariance function $R(v) = E\Big(Y(t)\ Y(t+v)\Big)$, or the variance $R(0)$ and the correlation function

$$\rho(v) = \frac{R(v)}{R(0)} = \mathrm{Corr}\Big(Y(t), Y(t+v)\Big) \quad .$$

The covariance function has a basic mathematical property called positive definiteness and defined as follows:

$$\sum_{j,k=1}^{n} c_j c_k^* R(v_j - v_k) \geq 0$$

for any integer $n$, complex numbers $c_1, \ldots, c_n$, and indices $v_1, \ldots, v_n$.

This property implies that $R(v)$ has a spectral representation which we write

$$R(v) = \int_0^1 e^{2\pi iuv} d\bar{F}(u) \quad ,$$

where $\bar{F}(u)$ is a non-decreasing function with $F(0) = 0$ . $\rho(v)$ has a spectral representation which we write

$$\rho(v) = \int_0^1 e^{2\pi iuv} dF(u)$$

where $F(u)$ is a distribution function (a non-decreasing function with $F(0) = 0$ and $F(1) = 1$ ). We call $F(\cdot)$ the spectral distribution function.

In developing the statistical theory of stationary time series analysis we always make the assumption that $\sum_v |\rho(v)| < \infty$ . Then the derivative $f(u) = F'(u)$ exists and is called the spectral density function; in terms of $f(\cdot)$ we have the spectral representations

$$\rho(v) = \int_0^1 e^{2\pi iuv} f(u)\, du \quad , \quad v = 0, \pm 1, \pm 2, \ldots, \; .$$

$$f(u) = \sum_{v=-\infty}^{\infty} e^{-2\pi iuv} \rho(v) \quad , \quad 0 \le u \le 1 \quad .$$

Our notation should be noted; we use $t$ to denote "time," $v$ to denote "lag" between two times, and $u$ to denote "frequency" when its domain is $0 \le u \le 1$ ; when frequency has other intervals in which it varies it

is customarily denoted by letters such as $\omega$ and $f$ and the intervals are $-\pi \leq \omega \leq \pi$ and $-0.5 \leq f \leq 0.5$. Note that, for a real valued time series, $f(u) = f(1 - u)$ and $\rho(v) = \rho(-v)$.

The mathematical existence of $f(u)$ is deduced from the fact that $\rho(v)$ is an integrable positive-definite function; the interpretation of $f(u)$ is deduced from the theory of linear filters.

To transform a stationary time series $Y(\cdot)$ to a new stationary time series $Z(\cdot)$, one generally uses linear time-invariant transformations (called filters) of the form

$$Z(t) = \sum_{j=0}^{\infty} b_j Y(t - j) \quad .$$

We like to introduce an operator (call it $B$ since its coefficients have been denoted $b_j$) such that one can write $Z(\cdot) = BY(\cdot)$. Define an operator $L$ (called the lag operator or backward shift operator):

$$Z(\cdot) = LY(\cdot) \text{ iff } Z(t) = Y(t - 1)$$

or equivalently $LY(t) = Y(t - 1)$; note $L^2Y(t) = Y(t - 2)$ and in general $L^nY(t) = Y(t - n)$ for any integer $n$. Introduce the power series

$$B(z) = \sum_{j=0}^{\infty} b_j z^j \quad ;$$

Then we can write $B = B(L)$ and $Z(t) = B(L)\, Y(t)$. We call $B(z)$ the transfer function of the filter $B(L)$. Regarded as a function of $z = e^{2\pi i u}$,

we call $B(e^{2\pi iu})$ the frequency response function. The notation has now been introduced to answer a basic question of stationary time series modeling: What are the properties of a time series $Z(\cdot)$ which arises as the output of a linear filter $B(L)$ whose input is a stationary time series $Y(\cdot)$.

Theorem. If $Z(t) = B(L)\, Y(t)$, where $Y(\cdot)$ is a zero mean stationary normal time series with spectral density function $f_Y(u)$, $0 \le u \le 1$, then $Z(\cdot)$ is a zero mean stationary normal time series with spectral density function $f_Z(u)$, $0 \le u \le 1$, given by

$$f_Z(u) = \left|B(e^{2\pi iu})\right|^2 f_Y(u) \quad .$$

Since many questions about a stationary time series $Y(\cdot)$ can be readily answered in terms of its spectral density function $f(u)$, it is natural that the estimation of $f(u)$ from a finite sample $\{Y(t),\ t = 1,\ldots,T\}$ should be one of the central problems of the theory of time series analysis.

Natural raw estimators $\tilde{\rho}$ and $\tilde{f}$ are obtained as follows: for $v = 0,1,\ldots,T-1$

$$\tilde{\rho}(v) = \sum_{t=1}^{T} Y(t)\, Y(t+v) \div \sum_{t=1}^{T} Y^2(t)$$

while $\tilde{\rho}(v) = 0$ for $v \ge T$ and $\tilde{\rho}(-v) = \tilde{\rho}(v)$; one may show that

$$\tilde{\rho}(v) = \int_0^1 e^{2\pi iuv}\, \tilde{f}(u)\, du \ ,$$

where

$$\tilde{f}(u) = \sum_{|v|<T} e^{-2\pi i u v} \tilde{\rho}(v)$$

$$= \left| \sum_{t=1}^{T} Y(t)\, e^{2\pi i t u} \right|^2 \Big/ \sum_{t=1}^{T} Y^2(t) \quad .$$

The convergence properties (as $T \to \infty$) of these estimators are as follows: $\tilde{\rho}(v) \to \rho(v)$ but $\tilde{f}(u) \not\to f(u)$. Indeed, $\tilde{f}(u)$ is in practice a very wiggly function which behaves like white noise in the sense that $\tilde{f}(u_1)$ and $\tilde{f}(u_2)$ are asymptotically independent for any fixed $u_1 \neq u_2$. The distribution of $\tilde{f}(u)$ is asymptotically exponential with mean $f(u)$. This is the point at which the modern era of time series analysis started (see, for example, Tukey (1959)): how to pass from wiggly estimators $\tilde{f}(u)$ to smooth estimators $\hat{f}(u)$ which are consistent (and, if possible, asymptotically "efficient") estimators of $f(u)$. In practice one might use and compare several estimators $\hat{f}(u)$ formed from the single finite sample of observations.

Three main approaches have developed for forming smooth estimators which are called the direct approach, the indirect approach, and the auto-regressive approach.

Each approach considers estimators or smoothers $\hat{f}(u)$ of a different form:

(i) Direct approach

$$\hat{f}(u) = \int_0^1 K(u - s)\, \tilde{f}(s)\, ds$$

for suitable kernels $K$.

(ii) Indirect approach

$$\hat{f}(u) = \sum_{v=-\infty}^{\infty} e^{-2\pi i u v} k(v) \tilde{\rho}(v)$$

for suitable weights $k$,

(iii) Autoregressive approach

$$\hat{f}(u) = \sigma_m^2 \left| 1 + \alpha_1 e^{2\pi i u} + \ldots + \alpha_m e^{2\pi i u m} \right|^{-2}$$

for a suitable integer $m$ (called the order), and coefficients $\sigma_m, \alpha_1, \ldots, \alpha_m$ which are estimated from the sample.

The extensive literature available on the properties of these methods of estimating $f(u)$ enables us to claim that we have successfully shown how to transform diverse statistical problems to a problem (density estimation) which has been "successfully" solved.

However, I would like to add a further claim; one can develop the autoregressive method so that it provides a "most successful" or "optimum" solution of the density estimation problem.

The name autoregressive approach comes from the notion of an autoregressive scheme. One can show that the true spectral density $f(u)$ is of the form

$$f(u) = \sigma_m^2 \left| g_m(e^{2\pi i u}) \right|^{-2}$$

where

$$g_m(z) = 1 + \alpha_1 z + \ldots + \alpha_m z^m$$

iff $\rho(v)$ satisfies the difference equation

$$\rho(v) + \alpha_1 \rho(1 - v) + \ldots + \alpha_m \rho(m - v) = 0 , \quad v > 0$$

$$= \sigma_m^2 , \quad v = 0 ,$$

iff $Y(t)$ satisfies the stochastic difference equation

$$Y(t) + \alpha_1 Y(t - 1) + \ldots + \alpha_m Y(t - m) = \epsilon(t)$$

where the process $\epsilon(t)$ obeys the conditions

$$E|\epsilon(t)|^2 = \sigma_m^2$$

$$\rho_\epsilon(v) = E\Big(\epsilon(t)\ \epsilon(t + v)\Big) = 0 \quad \text{for} \quad v \neq 0$$

$$E\Big(Y(s)\ \epsilon(t)\Big) = 0 \quad \text{for} \quad s < t$$

A time series is called <u>white noise</u> iff its correlation function $\rho(v) = 0$ for $v \neq 0$ .

Modeling a time series by an autoregressive scheme is convenient

because one can then: (i) readily estimate the parameters of the model, and (ii) solve the prediction problem: given the values $Y(t),Y(t-1),\ldots$ to predict $Y(t+1),Y(t+2),\ldots$ .

To a general spectral density $f(u)$ satisfying the conditions that $\log f(u)$ and $f^{-1}(u)$ are integrable we can associate a sequence of autoregressive approximators $f_m(u)$ , $m = 0,1,\ldots$ First, $f_0(u) = 1$ ; to define $f_m(u)$ for $m > 0$ introduce the minimization problem: let $\alpha_{1,m},\ldots,\alpha_{m,m}$ be the values at which

$$J_m(a_1,\ldots,a_m) = \int_0^1 \left|1 + a_1 e^{2\pi iu} + \ldots + a_m e^{2\pi ium}\right|^2 f(u)\, du$$

achieves its minimum value, and let $\sigma_m^2$ denote the minimum value so that

$$\sigma_m^2 = J_m\left(\alpha_{1,m},\ldots,\alpha_{m,m}\right)$$

Define

$$g_m(z) = 1 + \alpha_{1,m} z + \ldots + \alpha_{m,m} z^m \quad .$$

The coefficients in $g_m(z)$ can be determined from the normal equations

$$\int_0^1 g_m(e^{2\pi iu})\, e^{-2\pi iuv} f(u)\, du = 0 \quad , \qquad v = 1,\ldots,m$$

which is equivalent to

$$\rho(-v) + \alpha_{1,m}\rho(1-v) + \ldots + \alpha_{m,m}\rho(m-v) = 0$$

and

$$\sigma_m^2 = \int_0^1 |g_m(e^{2\pi i u})|^2 f(u)\, du$$

$$= \int_0^1 g_m(e^{2\pi i u})\, f(u)\, du$$

$$= \rho(0) + \alpha_{1,m}\rho(1) + \ldots + \alpha_{m,m}\rho(m)$$

Conditions for the convergence of $f_m(u)$ to $f(u)$ are stated in Geronimus (1960); in addition to $\log f(u)$ and $f^{-1}(u)$ are both integrable we must assume a certain sequence of partial correlation coefficients is absolutely summable. One can then show that one can represent

$$f(u) = \sigma_\infty^2 |g_\infty(e^{2\pi i u})|^{-2}$$

$$g_\infty(z) = 1 + \alpha_{1,\infty} z + \ldots + \alpha_{m,\infty} z^m + \ldots \quad .$$

Estimators $\hat{f}_m(u)$ of $f(u)$ are easily obtained as follows. Let $\hat{\alpha}_1,\ldots,\hat{\alpha}_m$ be the solutions of the sample normal equations

$$\tilde{\rho}(-v) + \hat{\alpha}_1\tilde{\rho}(1-v) + \ldots + \hat{\alpha}_m\tilde{\rho}(m-v) = 0 , \qquad v = 1,\ldots,m$$

Define

$$\hat{g}_m(z) = 1 + \hat{\alpha}_1 z + \ldots + \hat{\alpha}_m z^m$$

$$\hat{\sigma}_m^2 = \tilde{\rho}(0) + \hat{\alpha}_1 \tilde{\rho}(1) + \ldots + \hat{\alpha}_m \tilde{\rho}(m)$$

$$\hat{f}_m(u) = \hat{\sigma}_m^2 \left| \hat{g}_m(e^{2\pi i u}) \right|^{-2}$$

The functions $\hat{f}_0(u), \ldots, \hat{f}_{T-1}(u)$ can be regarded as a sequence of functions which proceed from the smoothest constant function $\hat{f}_0(u) = 1$ to the wiggliest function $\hat{f}_{T-1}(u) = \tilde{f}(u)$. One desires to find an intermediate value of $m$, denoted $\hat{m}$, such that $\hat{f}_{\hat{m}}(u)$ can be regarded as not the smoothest estimator of $f(u)$ but as the "most likely" estimator of $f(u)$. For this purpose one needs a criterion to determine $\hat{m}$ (called an order-determination criterion). Such criteria have been developed by a number of authors using various conceptual frameworks. The approach of Akaike (1974) is particularly well known. The discussion of this question requires an extensive paper by itself. Space permits me only to introduce my own criterion, which I call CAT (criterion autoregressive transfer function).

Rather than directly examining the properties of $\hat{f}_m(u)$ as an estimator of $f(u)$, I focus on the properties of $\hat{g}_m(z)$ as an estimator of $g_\infty(z)$. We would like to choose $\hat{g}_\infty(z)$ to minimize the overall mean square error

$$J = \int_0^1 E\left|\hat{g}_\infty(e^{2\pi iu}) - g_\infty(e^{2\pi iu})\right|^2 f(u)\, du$$

$$= \sigma_\infty^2 \int_0^1 E\left|\frac{\hat{g}_\infty - g_\infty}{g_\infty}\right|^2 du \quad .$$

Now overall mean square error can be expressed as a sum of overall variance

$$V = \int_0^1 \mathrm{Var}\,[\hat{g}_\infty]\, f(u)\, du$$

and overall squared bias

$$B^2 = \int_0^1 \left|E\hat{g}_\infty - g_\infty\right|^2 f(u)\, du$$

It can be shown (see Parzen (1976)) that the degree $m$ polynomial best approximating $g_\infty$ is $g_m$ multiplied by a suitable constant. Therefore we restrict ourselves to estimators $\hat{g}_\infty$ of the form $\hat{g}_\infty = \hat{g}_{\hat{m}}$ where $\hat{m}$ minimizes the function of $m$

$$J(m) = \int_0^1 E\left|\hat{g}_m - \hat{g}_\infty\right|^2 f(u)\, du$$

$$= \int_0^1 \mathrm{Var}\,(\hat{g}_m)\, f(u)\, du + \int_0^1 \left|g_m - g_\infty\right|^2 f(u)\, du$$

We are able to obtain a remarkable approximate evaluation of $J(m)$ by changing our definitions. Define

$$\gamma_\infty(z) = \frac{g_\infty(z)}{\sigma_\infty^2} , \quad \gamma_m(z) = \frac{g_m(z)}{\sigma_m^2} , \quad \hat{\gamma}_m(z) = \frac{\hat{g}_m(z)}{\hat{\sigma}_m^2}$$

and change the definition of $J(m)$ to

$$J(m) = \int_0^1 \text{Var}\ (\hat{\gamma}_m)\ f(u)\ du + \int_0^1 |\gamma_\infty - \gamma_m|^2\ f(u)\ du$$

One can show that the second term (representing the overall bias) equals $\sigma_\infty^{-2} - \sigma_m^{-2}$, and that

$$J(m) \doteq \frac{1}{T} \sum_{j=1}^{m} \sigma_j^{-2} + \sigma_\infty^{-2} - \sigma_m^{-2} .$$

This remarkable formula motivates the following order determination criterion: given a sample of size $T$, choose $\hat{m}$ to minimize the function $CAT(m)$ calculated from the sample as follows:

$$CAT(0) = -\left(1 + \frac{1}{T}\right) .$$

$$CAT(m) = \frac{1}{T} \sum_{j=1}^{m} \hat{\hat{\sigma}}_j^{-2} - \hat{\hat{\sigma}}_m^{-2}$$

where $\hat{\hat{\sigma}}_m^2$ is an "unbiased" estimator of $\sigma_m^2$ defined by

$$\hat{\hat{\sigma}}_m^2 = \left(1 - \frac{m}{T}\right)^{-1} \hat{\sigma}_m^2 .$$

When $\hat{m} = 0$, we estimate $f(u)$ to be the constant 1, and accept

the hypothesis that the sample could have been drawn from a white noise process.

An order determination procedure can be regarded as a procedure for adaptively determining from the sample a "most powerful" test statistic for the null hypothesis of white noise. The meaning of this assertion requires another paper to discuss.

It may be helpful to make some intuitive remarks about order-determining criteria. The residual variances $\hat{\sigma}_m$ decrease as $m$ increases so that they do not decisively indicate which order $\hat{m}$ is long enough. The minimum of the "unbiased residual variances" $\hat{\hat{\sigma}}^2_m$ usually exists; while empirically it may on occassion choose the "right" order there is no conceptual basis for its use. Akaike's criterion, to minimize

$$AIC(m) = \log \hat{\sigma}^2_m + 2\frac{m}{T} ,$$

can be justified using an entropy maximization inference criterion. In recent work, Wahba uses cross-validation inference criteria to determine smoothing factors; her work can be directly applied to density estimation.

I would like to suggest a new criterion, motivated by the cross-validation criteria of Wahba and which I call CV, whose order-determining properties need to be examined:

$$CV(m) = \frac{\hat{\sigma}^2_m}{\left\{\int_0^1 |\hat{g}_m(e^{2\pi i u})|\, du\right\}^2} .$$

To estimate a density function $d(u_1,..,u_m)$ which is a function of several variables, one cannot use autoregressive methods of approximation; however, one can develop indirect methods of estimation where the weights are chosen using cross-validation criteria. I believe we are justified in claiming that one can empirically estimate density functions with almost no prior assumptions.

## 5. Reliability Theory

Let $f_0Q_0(u)$ be the density quantile function of the exponential distribution $f_0(x) = e^{-x}$ ; then $Q_0(u) = -\log(1-u)$ and $f_0Q_0(u) = 1-u$ . Let $f(x)$ be the probability density of a non-negative random variable $X$ ; then

$$\int_0^1 \frac{1-u}{fQ(u)}\,du = \int_0^1 (1-u)\,q(u)\,du = \int_0^\infty \{1-F(x)\}\,dx = \mu = E[X] \quad .$$

Thus integrability of $(1-u)\,q(u)$ is equivalent to the mean being finite. The integrand $(1-u)\,q(u)$ occurs frequently as it is related to the hazard function

$$h(x) = \frac{f(x)}{1-F(x)}$$

and the hazard quantile function

$$hQ(u) = h\left(F^{-1}(u)\right) = \frac{fQ(u)}{1-u} = \frac{1}{(1-u)\,q(u)}$$

Next define the distribution function

$$F_{res}(x) = \int_0^x \frac{1}{\mu} \{1 - F(y)\}\, dy$$

which is denoted $F_{res}$ as it is the distribution function of the residual lifetime in a renewal process. The distribution function on $0 \le u \le 1$

$$D(u) = F_{res} F^{-1}(u)$$

$$= \frac{1}{\mu} \int_0^{Q(u)} \{1 - F(y)\}\, dy$$

$$= \int_0^u (1 - t)\, q(t)\, dt \left\{ \int_0^1 (1 - t)\, q(t)\, dt \right\}^{-1}$$

has density $d(u) = \frac{1}{\mu} \frac{(1 - u)}{fQ(u)}$ . The following hypotheses are equivalent:

$D(u) = u$ ,

$d(u) = 1$ ,

$F_{res}(x) = F(x)$

$F(x)$ is the exponential distribution

In other words, a <u>test for exponentially</u> is providing by testing whether the density function $d(u)$ is constant.

A raw estimator of $D(u)$ is provided by

$$\tilde{D}(u) = \frac{\int_0^u (1 - s)\, dQ_n(s)}{\int_0^1 (1 - s)\, dQ_n(s)}$$

This function has been extensively studied by researchers in reliability theory (especially Barlow and Van Zwet (1970), Barlow and Proschan (1976)) under the name of the total time on test statistic.

The statistic $\tilde{D}(u)$, $0 \leq u \leq 1$, also can be deduced from the general one-sample theory outlined in Section 1; however, the derivation is different in this section since it is directly motivated by a search for a test of exponentiality. Section 1 provides tests for any specified distribution (more precisely, a specified density-quantile or $f_0 Q_0$ function). A list of density-quantile functions of familiar univariate distributions is given in Table I.

Table I Density-Quantile Functions

| Name of Probability Law | Density $f(x)$ | Quantile $Q(u)$ | Density-Quantile $fQ(u)$ |
|---|---|---|---|
| Normal | $\varphi(x) = \Phi^{1}(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ | $\Phi^{-1}(u)$ | $\frac{1}{\sqrt{2\pi}} \exp -\frac{1}{2}\left\|\Phi^{-1}(u)\right\|^2$ |
| Log-normal | $\frac{1}{x}\varphi(\log x)$ | $e^{\Phi^{-1}(u)}$ | $\varphi\Phi^{-1}(u)\, e^{-\Phi^{-1}(u)}$ |
| Exponential | $e^{-x}$, $x > 0$ | $-\log (1 - u)$ | $1 - u$ |
| Pareto $\beta > 0$ | $\{\beta x^{1+(1/\beta)}\}^{-1}$, $x > 1$ | $(1 - u)^{-\beta}$ | $\frac{1}{\beta}(1 - u)^{1+\beta}$ |
| Extreme Value | $e^{x} e^{-e^{x}}$ | $\log \log (1 - u)^{-1}$ | $(1 - u) \log \frac{1}{1 - u}$ |
| Weibull $c = 1/\beta > 0$ | $cx^{c-1} e^{-x^{c}}$, $x > 0$ | $\left\{\log \frac{1}{1 - u}\right\}^{\beta}$ | $\frac{1}{\beta}(1 - u)\left\{\log \frac{1}{1 - u}\right\}^{1-\beta}$ |
| Cauchy | $\frac{1}{\pi} \frac{1}{1 + x^2}$ | $\tan \pi(u - \frac{1}{2})$ | $\frac{1}{\pi} \sin^2 \pi u$ |
| Logistic | $\frac{e^{x}}{(1 + e^{x})^2}$ | $\log \frac{u}{1 - u}$ | $u(1 - u)$ |
| Double-exponential | $\frac{1}{2} e^{-\|x\|}$ | $\log 2u$, $u < \frac{1}{2}$<br>$-\log 2(1 - u)$, $u > \frac{1}{2}$ | $u$, $u < \frac{1}{2}$<br>$1 - u$, $u > \frac{1}{2}$ |
| Uniform-reciprocal | $\frac{1}{x^2}$, $x > 1$ | $\frac{1}{1 - u}$ | $(1 - u)^2$ |

Score functions J(u)

| | |
|---|---|
| Normal | $\Phi^{-1}(u)$ |
| Exponential | 1 |
| Double-Exponential | sign (2u - 1) |
| Logistic | 2 u - 1 |
| Cauchy | - sin $2\pi u$ |

## References


Akaike, H. (1974) A new look at the Statistical Model Identification. IEEE Trans. Automat. Control, AC-19, 716-123.

Akaike, H. (1976) On Entropy Maximization Principle. Symposium on Applications of Statistics to be edited by P. Krishnaiah.

Barlow, R. and Proschan, F. (1977) Asymptotic Theory of Total Time on Test Processes, with Applications to Life Testing. Multivariate Analysis IV edited by P. Krishnaiah, Academic Press: New York.

Barlow, R. E. and Van Zwet, W. R. (1970) Asymptotic Properties of Isotonic Estimators for the Generalized Failure Rate Function. Part I: Strong Consistency, in M. L. Puri, ed., Nonparametric Techniques in Statistical Inference, Cambridge Univ. Press, 159-173.

Chernoff, H. and Savage, I. R. (1958) Asymptotic normality and efficiency of certain nonparametric test statistics, Ann. Math. Statist. 29, 972-994.

Easterling, R. G. (1976) Goodness of Fit and Parameter Estimation, Technometrics, 18, 1-9.

Geronimus, Y. L. (1960) Polynomials Orthogonal on a Circle and Interval (translated from Russian), Pergamon: New York.

Hajek, J. (1969) Nonparametric Statistics. Holden Day: San Francisco.

Hamdan, M. A., Pirie, W. R. and Khuri, A. I. (1976) Unbiased estimation of the Common Mean Based on Incomplete Bivariate Normal Samples, Biometrische Zeitschrift, 18, 245-249.

Kimeldorf, G. and Sampson, A. (1975) Uniform Representations of Bivariate Distributions, Comm. in Stat. 4, 617-627.

Kullback, S. (1958) Information Theory and Statistics, Wiley: New York

Parzen, E. (1961) Regression analysis of continuous parameter time series, Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Univ. California Press, Berkeley, Calif., Vol. I, 469-489.

Parzen, E. (1970) Statistical inference on time series by RKHS methods, II, Proceedings of the 12th Bienniel Canadian Mathematical Congress, edited by R. Pyke, American Mathematical Society, Providence, R. I., 1-37.

Parzen, E. (1974) Some Recent Advances in Time Series Modeling, IEEE Transactions on Automatic Control, AC-19, 723-730.

Parzen, E. (1975) Some Solutions to the Time Series Modeling and Prediction Problem, The Search for Oil, edited by D. Owens. Marcel Dekker: New York, 1-16.

Pyke, R. (1965) Spacings (with discussion). J. Roy. Statist. Soc. B, 27, 395-449.

Pyke, R. (1972) Spacings Revisited, Proceedings Sixth Berkeley Symposium on Mathematical Statistics and Probability, ed. L. LeCam and J. Neyman, University of California Press, I, 417-427.

Pyke, R. and Shorack, G. (1968) Weak convergence of a two-sample empirical process and a new proof to the Chernoff-Savage theorem. Ann. Math. Statist., 39, 755-771.

Ruymgaart, F. H. (1974) Asymptotic normality of non-parametric tests for independence. Ann. Statistics, 2, 892-910.

Shorack, G. (1972) Convergence of Quantile and Spacings Processes with Applications. Ann. Math. Statist., 43, 1400-1411.

Tukey, J. W. (1959) An introduction to the measurement of spectra, Probability and Statistics, edited by U. Grenander, Wiley: New York, 300-330.

Wahba, G. (1976) A survey of some smoothing problems and the method of generalized cross-validation for solving them. Symposium on Applications of Statistics to be edited by P. Krishnaiah.

Weiss, L. (1964) On the Asymptotic Joint Normality of Quantiles from a Multivariate Distribution. Jour. Research Nat. Bur. Standards B, 68, 65-66.

Weiss, L. and Wolffowitz, J. (1970) Asymptotically Efficient Non-Parametric Estimators of Location and Scale Parameters, Z. Wahrscheinlichkeits-Theorie Revw. Gebiete, 16, 134-150.